42390.P11905 PATENT

UNITED STATES PATENT APPLICATION

for

A SCALABLE CODING SCHEME FOR LOW LATENCY APPLICATIONS

Applicants:

Wen-Hsiao Peng
Yen-Kuang Chen

prepared by:

BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN
12400 Wilshire Boulevard
Los Angeles, CA 90026-1026
(408) 720-8598

EXPRESS MAIL CERTIFICATE OF MAILING

"Express Mail" mailing label number EL627534345US
Date of Deposit September 26, 2001
I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Commissioner of Patents and Trademarks, Washington, D.C. 20231.

Michelle Begay
(Typed or printed name of person mailing paper or fee)

Michelle Begay
(Signature of person mailing paper or fee)

A SCALABLE CODING SCHEME FOR LOW LATENCY APPLICATIONS

FIELD OF THE INVENTION

This invention relates generally to video encoding and decoding, and more particularly to a scalable coding scheme for video encoding and decoding.

BACKGROUND OF THE INVENTION

Video is principally a series of still pictures, one shown after another in rapid succession, to give a viewer the illusion of motion. Before it can be transmitted over a communication channel, analog video may need to be converted, or "encoded," into a digital form. In digital form, the video data are made up of a series of bits called a "bitstream." When the bitstream arrives at the receiving location, the video data are "decoded," that is, converted back to a viewable form. Due to bandwidth constraints of communication channels, video data are often "compressed" prior to transmission on a communication channel. Compression may result in a degradation of picture quality at the receiving end.

A compression technique that partially compensates for loss (degradation) of quality involves separating the video data into a "base layer" and one or more "enhancement layers" prior to transmission. The base layer includes a rough version of the video sequence and may be transmitted using comparatively little bandwidth. The enhancement layers typically capture the difference between the base layer and the original input video picture. Each enhancement layer also requires little bandwidth, and one or more enhancement layers may be transmitted at the same time as the base layer.

At the receiving end, the base layer may be recombined with the enhancement layers during the decoding process. The enhancement layers provide correction to the base layer, consequently improving the quality of the output video. Transmitting more enhancement layers produces better output video, but requires more bandwidth.

The enhancement layers may be ordered so that the most significant correction is made by the first enhancement layer, with subsequent enhancement layers providing less significant correction. In this way, the quality of the output video can be "scaled" by combining different numbers of the ordered enhancement layers with the base layer. The process of using ordered enhancement layers to scale the quality of the output video is referred to as "Fine Granularity Scalability" (FGS) and may result in a substantial saving of bandwidth.

Some compression methods and file formats have been standardized, such as the Motion Picture Experts Group (MPEG) standards of the International Organization for Standardization. One of the MPEG standards, MPEG-4, uses an FGS algorithm to produce a range of quality of output video suitable for use with various bandwidths. However, the amount of processing required with MPEG-4 FGS renders it unsuitable for applications that require low end-to-end delay, such as videoconferencing.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1A is a block diagram illustrating a video conferencing system in accordance with the present invention;

Figure 1B is a block diagram illustrating a video streaming system in accordance with the present invention;

Figure 2 is a block diagram illustrating a prior art encoding structure;

Figure 3 is a block diagram illustrating one embodiment of an encoding structure according to the invention;

Figure 4 is a block diagram illustrating a prior art decoding structure corresponding to the encoding structure of Figure 2;

Figure 5 is a block diagram illustrating a decoding structure corresponding to the encoding structure of Figure 3;

Figures 6A-C are block diagram illustrating alternate embodiments of a decoding structure according to the invention;

Figures 7A-C are block diagrams illustrating encoding structures corresponding to the decoding structures of Figures 6A-C;

Figure 8 is a flowchart of an method for performing the encoding operation of the encoding structures of Figures 7A-C;

Figures 9A-C are flowcharts of methods for performing the decoding operations of decoding structures of Figure 6A-C; and

Figure 10 is a diagram of one embodiment of a computer system in which a encoder or decoder according to the invention may incorporated.

DETAILED DESCRIPTION OF THE INVENTION

In the following detailed description of embodiments of the invention, reference is made to the accompanying drawings in which like references indicate similar elements, and in which is shown by way of illustration specific embodiments in which the invention may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention, and it is to be understood that other embodiments may be utilized and that logical, mechanical, electrical, functional

and other changes may be made without departing from the scope of the present invention. The following detailed description is, therefore, not to be taken in a limiting sense, and the scope of the present invention is defined only by the appended claims.

Two video systems in which embodiment of the invention may be practiced are shown in Figures 1A and 1B. Figure 1A is a block diagram of a video conferencing system 100 in which the participant's video system 101, 103 contain an MPEG-4 FGS codec 105 that encodes and decodes video streams in accordance with embodiments of the invention described below. Video system 101 is connected to video system 103 through an asymmetric communications link that has more bandwidth in the channel 119 from video system 101 to video system 103 than the channel 125 from video system 103 to video system 101, such as an asymmetric digital subscriber line (ADSL) or digital cable connections. As shown, an encoder 107 in the codec 105 at video system 101 encodes a video signal from a camera 111 into a base layer bitstream and an enhancement layer bitstream and sends the bitstreams to a communications interface 115 for transmission to the video system 103. The communications interface 115 transmits the largest amount of bitstream data that can be handled by the channel 119. The bitstreams may be combined into a single bitstream through a multiplexer (not shown) before they are transmitted. When a communications interface 117 at video system 103 receives the bitstream, the bitstream is de-muxed, if necessary, and the base and enhancement layer bitstream are sent to a decoder 109 in the codec 105, where they are decoded into a video picture that is shown on a display 121. Similarly, the encoder 107 in the code 105 at video system 103 encodes a video signal from camera 123. However, because the channel 125 is of lower bandwidth than the channel 119, a communications

interface 117 on video system 103 selects less bitstream data to transmit to video system 101.

In a heterogeneous networking environment the bandwidth of the channels can vary significantly but the scalability of MPEG-4 FGS enables the video systems 101, 103 to transmit the highest quality video given the available bandwidth. However, in a real-time video system, such as the video conferencing system of Figure 1A, a low end-to-end latency cannot be achieved if the codec must perform processing-intensive calculations on each end of the transmission. As described below, in one embodiment, the present invention reduces the amount of processing required to encode the video stream by using a fractional part of existing quantized video coefficients as frequency-ordered enhancement layers to eliminate a separate frequency weighting component traditionally used to reduce flickering. For environments in which motion compensation is not critical, such as video conferences, the use of the fractional part as the frequency-ordered enhancement layers also allows the reuse of decoding components.

The embodiments of the present invention are not limited to use with low latency video systems but is equally applicable to streaming video systems, such as illustrated in Figure 1B. An encoder 133 in an encoding system (not shown) encodes a video signal from a camera 131 into base and enhancement layer bitstreams, which are stored on a storage device 135. When a user requests the video, a server 137 reads the bitstreams from the storage device 135, determines the bandwidth of the communications link 139 to the playback computer 141, and transmits an amount of bitstream data that is supported by the bandwidth. A decoder 143 in the playback system 141 decodes the bitstream to produce the video shown to the user on display 145. Incorporation of the invention reduces the processing requirements of the encoder 133 and decoder 143 and

thus that of both the encoding and the playback systems when producing and displaying high quality video.

While various system configurations have been described to illustrate the use of encoder and decoders, it will be appreciated that the invention is not limited to these configurations. For example, the encoding system and the server 137 of Figure 1B may be the same or different systems. Furthermore, any or all of the video systems, the encoding system, the server, and the playback system may be general purpose computers, such as described below in conjunction with Figure 10, or specially-designed systems.

The use of the fractional part of the quantization as the enhancement layer in an FGS codec is now described in conjunction with Figures 3 and 5, with reference to prior art encoding and decoding structures in Figures 2 and 4. The FGS encoding structure 200 of Figure 2 encodes a series of video frames 203 to produce a base layer bitstream 213 plus a bitstream of one or more enhancement layers 237. A base layer encoding structure 201 employs a discrete cosine transform (DCT) 207, quantization (Q) 209, and variable length coding (VLC) 211 components. The encoding structure also includes a feedback "reconstruction" loop that subtracts 205 a reconstructed base layer from an incoming frame 203 to remove temporal redundancies from the incoming frames. The reconstruction loop performs an inverse quantization (IQ) 215 and inverse discrete cosine transform (IDCT) 217 on the output of Q 209 to reverse the IQ 215 and DCT 217 operations. A clipping component 221 modifies the output of the IDCT component 217 to within a valid range, if needed, and the output from the clipping component 221 is stored in a frame memory 223. The data in the frame memory 223 is processed to

compensate for motion (using motion estimation (ME) 225 and motion compensation (MC) 227 components) to produce the reconstructed base layer.

An enhancement layer encoding structure 202 produces the enhancement layers by subtracting 229 the clipping component output from the incoming frame 203 and transforming the difference into coefficients in the DCT domain (DCT 231). When lower frequency layers are transmitted first, the flickering effect is reduced, and so a frequency weighting (FW) 233 shifts each DCT coefficient using a FW matrix to arrange the individual enhancement layers in frequency order. The ordered enhancement layers are processed through a VLC 235 to produce the enhancement layer bitstream 237.

It can be shown that the FW matrix is equivalent to a quantization matrix with a stepsize of a power of two and that a quantization matrix that satisfies the following equation is functionally equivalent to the FW matrix:

$$\frac{E(|\Delta x|)}{Qx} \geq \frac{E(|\Delta y|)}{Qy} \quad where \quad Qx \leq Qy \qquad (1)$$

with Q_x and Q_y being the quantization stepsize of the low and high frequency DCT coefficients, respectively, and Δx and Δy representing the remainder of low and high frequency DCT coefficients, respectively. Assuming a base layer quantization matrix having the above properties in the Q component 209, and the IDCT 217 and DCT 231 having infinite precision, the fractional part of the quantized base layer DCT coefficients is functionally equivalent to the output of the frequency weighting component 233.

While a quantization matrix has been described that creates a fractional part equivalent to frequency-ordered enhancement layers, one of skill in the art will immediately recognize that other embodiments of encoding structures may require

alternate quantization matrices that create fractional parts equivalent to enhancement layers ordered according to other criteria or enhancements layers that are not arranged in any particular order. Such alternate quantization matrices are considered within the scope of the invention.

Thus, the enhancement layer encoding structure 202 can be modified as illustrated in Figure 3. An appropriate quantization matrix is incorporated into Q 209 and the quantized base layer DCT coefficients are parsed into their integer parts 305 and fractional parts 307. The integer parts 305 are input into the VLC 211 to produce a base layer bitstream 309 and into the reconstruction loop. The fractional parts 307 are variable length encoded 235 in an enhancement layer encoding structure 303 to produce an enhancement layer bitstream 311. Each enhancement layer may be represented by one or more binary decimal positions within each fractional part 307. The fractional parts 307 may not be able to be represented by a finite number of binary digits and some bits may need to be truncated. Therefore, in one embodiment, a maximum number to keep is based on the capacity of the system on which the encoding structure is implemented. In an alternate embodiment, as many bits as possible are kept, which may vary from time to time depending on the load on the system. It will be appreciated that the stepsize of the quantization matrix may be any integer value N, and the resulting integer part will be a value from 0 to N-1.

A prior art decoding structure 400 corresponding to the prior art encoding structure 200 is illustrated in Figure 4. A base layer decoding structure 401 reverses the operations of the encoding structure 200 on the base layer bitstream 213 using a variable length decoder (VLD) 403, inverse quantization (IQ) 405 and an inverse discrete cosine transform (IDCT) 407. A feedback "prediction" loop adds 409 some or all of the

temporal redundancies removed in the reconstruction loop in the base layer encoding structure 201 to create a restored base layer. Clipping 411 is performed on the restored base layer and the result is (optionally) output as the base layer 413 and stored in a frame memory 415. A motion compensation (MC) component 417 is also included in the prediction loop. The enhancement bitstream 237 is decoded in structure 403 using a VLD 419, a bit plane shifter (BP shift) 421 to reverse the FW operation 233 and an IDCT 423. The base layer 413 is added 425 to the resulting enhancement layers, clipped 427, and output as the video frame 429.

Because the enhancement bitstream 311 produced by the encoder 300 in Figure 3 is derived from the fractional parts of the quantized DCT coefficients computed at the base layer, the enhancement layer decoding structure 403 can be modified as shown in Figure 5. After variable length decoding 419, an inverse quantization (IQ) component 507 is used to reverse the quantization 209 that produced the fractional parts. The IQ 507 replaces the BP shift component 421 shown in Figure 4.

It will be appreciated that corresponding DCT/IDCT, Q/IQ and VLC/VLD components are used within the encoding/decoding structures illustrated herein. Additionally, common components are given different reference numbers to indicate different instances of a component. For example, one of skill in the art will immediately recognize that IQ 507 and IQ 405 in Figure 5 perform the same operations but cannot be the same IQ component because the enhancement layer signal would flow into the prediction loop and result in error drifting.

When motion compensation for the video frames is not critical, the encoding and decoding structures of Figures 3 and 5 may be modified further as illustrated in Figures 6A-C and 7A-C, and the particular embodiments illustrated in Figures 6B and 6C allow

the reuse of common decoding components without causing error drifting. In each of the following embodiments, the decoding structure is first described and then the corresponding encoding structure. Clipping components are not shown for ease in illustration but one of skill in the art will immediately understand where clipping would be applied.

Assuming zero-motion vectors, the prediction loop in the base layer decoding structure 501 can be treated as a linear time invariant loop and does not require the MC component 417, resulting in the base layer decoding structure 601 illustrated Figure 6A. Furthermore, the addition component 425 can be moved into the base layer decoding structure 601, eliminating output of the base layer alone when there are enhancement layers present. The corresponding encoding structure 700 illustrated in Figure 7A eliminates the ME 225 and MC 227 components in the reconstruction loop since the reconstruction loop is similarly treated as a linear time invariant loop.

In another embodiment, the exchange law of linear time invariant systems allows the prediction loop in the base layer decoding structure to be exchanged with the IDCT 407 as illustrated in Figure 6B. Because the IDCT 407 and the IDCT 423 of Figure 6A are equivalent, exchanging the prediction loop with the IDCT 407 allows the use of a single IDCT without incurring error drifting problems. The corresponding encoding structure 710 is shown in Figure 7B, in which the temporal redundancies are removed from the incoming frames 203 after the frames have been transformed into the DCT domain.

In a further embodiment, the prediction loop is moved before the IQ 405 as illustrated in Figure 6C to allow reuse of the IQ 405 in the decoding structure 620 without introducing error drifting. Because the IQ 405 is not a linear time invariant

system, the quantization parameters Q_p used to encode the bitstreams 725, 727 (Figure 7C) are transmitted to the decoding structure so the time variations of IQ 405 are known to the decoding structure 620. It will be appreciated that having Q_p fixed or smoothly changing will enhance the efficiency of the decoding structure 620. A corresponding encoding structure 720 illustrated in Figure 7C removes the temporal redundancies from the incoming frames 203 after the frames have been transformed and quantized.

Next, the particular methods of the invention are described in terms of executable instructions with reference to a series of flowcharts. Describing the methods by reference to a flowchart enables one skilled in the art to develop such instructions to carry out the methods within suitably configured processing units. The executable instructions may be written in a computer programming language or may be embodied in firmware logic. Furthermore, it is common in the art to speak of executable instructions as taking an action or causing a result. Such expressions are merely a shorthand way of saying that execution of the instructions by a computer causes the processor of the computer to perform an action or a produce a result.

Figures 8 and 9A-C illustrate the methods to be performed by a system to implement the operations of the encoding and decoding structures of the invention previously described. The embodiments of an encoding method are illustrated in a flowchart in Figure 8, including all the acts (blocks) from 801 until 819. The embodiments of decoding methods are illustrated in flowcharts in Figures 9A-C, including all the acts from 901 until 935.

Referring first to Figure 8, the acts that cause a system to perform the operations of the encoding structures of Figures 3 and 7A-C are shown as FGS encoding method 800. Because the encoding structures of Figures 3 and 7A-C remove the temporal

redundancies at different points in the encoding process, phantom blocks 801, 805 and 815 in Figure 8 are used to represent the removal operation. For Figures 3 and 7A, temporal redundancy removal is performed at block 801; for Figure 7B, at block 805; and for Figure 7C, at block 815. The incoming frame is transformed (block 803) and the transformed result is quantized (block 807). The result of the quantization is parsed at block 809 into its fractional parts and its integer parts. The fractional parts are encoded (block 811), and the encoded fractional parts are output as the enhancement layer bitstream (block 813). The integer parts are encoded (block 817), and the encoded integer parts are output as the base layer bitstream (block 819).

An FGS decode method 900 is illustrated in Figure 9A that causes a system to perform the operations of the decode structures shown in Figures 5 and 6A. The enhancement layers are decoded (block 901), an inverse quantization (block 903) and an inverse transformation (block 905) are applied. The output of block 905 is applied into the base layer at block 915 as described below. Similarly, the base layer is decoded (block 907) and an inverse quantization (block 909) and an inverse transformation (block 911) applied. Some or all of the base layer temporal redundancies are restored to reconstruct the base layer (block 913). At block 915, the enhancement layers from block 905 are applied to the reconstructed base layer. The resulting video stream is output at block 917.

Turning now to Figure 9B, an FGS decode method 920 is described that causes a system to perform the operations of the decoding structures of Figure 6B. Similar to method 900, method 920 decodes the enhancement layer data stream at block 901 and applies an inverse quantization at block 903. The output of block 903 is applied to the base layer at block 915 as described below. As in Figure 9A, the base layer bitstream is

decoded at block 907 and the inverse quantization is applied at block 909. The base layer temporal redundancies are restored at block 913 and the enhancement layers from block 903 are applied to the modified base layer at block 915. In this case, the inverse transformation is applied to the combined base layer and enhancement layers (block 921) and the result is output as the video stream (block 923).

As shown in Figure 9C, an FGS decode method 930 that that causes a system to perform the operations of the decoding structures in Figures 6C decodes the enhancement layer data stream at block 909, leaving the inverse quantization and the inverse transformation to be applied to the combined base layer and enhancement layers at blocks 931 and 933, with the resulting video stream being output at 935. Blocks 907, 913 and 915 perform the processes described above for those blocks in Figures 9A and 9B.

The following description of Figure 10 is intended to provide an overview of computer hardware environments in which the encoding/decoding structures and methods of the invention can be implemented, but is not intended to limit the applicable environments. Figure 10 shows one example of a conventional computer system 1001 containing a processing unit 1005 and a memory 1009 coupled to the processor 1005 by a bus 1007. Memory 1009 can be dynamic random access memory (DRAM) and can also include static RAM (SRAM). The bus 1007 couples the processor 1005 to the memory 1009 and also to non-volatile storage 1015, to display controller 1011, to the input/output (I/O) controller 1017, and to a modem or network interface 1003. The display controller 1011 controls in the conventional manner a display on a display device 1013 which can be a cathode ray tube (CRT) or liquid crystal display. The input/output devices 1019 can include a keyboard, disk drives, printers, a scanner, and other input and

output devices, including a mouse or other pointing device. The input/output devices 1019 may include a digital image input device, such as a digital camera, that is coupled to the I/O controller 1017 in order to allow images from the digital image input device to be input into the computer system 1001. The modem/network interface 1003 enables the computer 1001 to communicate with other computers or devices on a network 10021. The display controller 1011, the I/O controller 1017, and the modem/network interface 1003 can be implemented with conventional well known technology. The non-volatile storage 1015 is often a magnetic hard disk, an optical disk, or another form of storage for large amounts of data. Some of this data is often written, by a direct memory access process, into memory 1009 during execution of software in the computer system 1001. One of skill in the art will immediately recognize that the term "computer-readable medium" or "machine-readable medium" includes any type of storage device that is accessible by the processor 1005 and also encompasses a carrier wave that encodes a data signal.

It will be appreciated that the computer system 1001 is one example of many possible computer systems which have different architectures. One of skill in the art will immediately appreciate that the invention can be practiced with other computer system configurations, including hand-held devices, multiprocessor systems, microprocessor-based or programmable consumer electronics, networked personal computers, minicomputers, mainframe computers, and the like. A typical computer system will usually include at least a processor, memory, and a bus coupling the memory to the processor.

Although specific embodiments have been illustrated and described herein, it will be appreciated by those of ordinary skill in the art that any arrangement which is

calculated to achieve the same purpose may be substituted for the specific embodiments shown. This application is intended to cover any adaptations or variations of the present invention. For example, although the invention is described in terms of particular FGS encoding and decoding structures, the concept of replacing the enhancement layer frequency weighting matrix with the base layer quantization matrix is applicable to other coding algorithms. Therefore, it is manifestly intended that this invention be limited only by the following claims and equivalents thereof.